DOCUMENT RESUME

ED 451 237 TM 032 465

AUTHOR Bastick, Tony
TITLE Accuracy of Same-Subject Estimates: Are Two Judgements Better Than One.
PUB DATE 1999-07-00
NOTE 8p.; Paper presented at the European Meeting of the Psychometric Society (11th, Luneburg, Germany, July 1999). For a related paper from the same author, see TM 032 461.
PUB TYPE Reports - Research (143) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Attention; *Cognitive Processes; *College Students; *Estimation (Mathematics); Foreign Countries; Higher Education; Timed Tests
IDENTIFIERS *Accuracy

ABSTRACT

The accuracy of the mean of two estimates was compared with the accuracy of a single independent estimate from the same subject. A subject was asked to estimate the size of one attribute of a constant stimulus, e.g., the total of a set of numbers. The same subject was also asked to give an estimate for an upper and lower bound on the size of the same attribute of the same stimulus. The experiment was designed to ensure that the single and double estimates were independent and given by the subjects under the same experimental conditions. The experimental design compared the accuracy of estimates of two stimulus attributes using a 3 by 4 randomized block design. Participants were 187 college students competing for $10 prizes with the time severely limited. Only 50 subjects managed to complete the tasks. The frequency with which one estimate was either extremely high or extremely low suggests that the levels of task complexity were too high for the stress level of the timed competition. The results have potential ramifications for methods of collecting judgmental data, but future research should use a task of more appropriate complexity. (Contains 1 figure, 2 tables, and 19 references.) (SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ED 451 237

Accuracy of Same-subject Estimates: Are Two Judgements Better than One

Author:
Tony Bastick

TM032465

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
✓ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Tony Bastick

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

BEST COPY AVAILABLE

TH

Accuracy of Same-Subject Estimates: Are Two Judgements Better Than One

Tony Bastick
University of the West Indies

Introduction

This research compares the accuracy of the mean of two estimates with the accuracy of a single independent estimate from the same subject. A subject was asked to estimate the size of one attribute of a constant stimulus e.g. the total of a set of numbers. The same subject was also asked to give an estimate for an upper and lower bound on the size of the same attribute of the same stimulus. The experiment was designed to ensure that the single and double estimates were independent and given by the subject under the same experimental conditions. The experimental design compared the accuracy of estimates of two stimulus attributes using a 3 by 4 randomised block design replicated over 50 subjects.

Traditional measurement and psychological theory predicts that the accuracy of a single estimate will be equal to the accuracy of the mean of a double estimate. A significant difference could challenge the statistical assumption traditionally applied in this type of application that the double estimate process is a simple replication of the single estimate process. The paper tests this assumption. The results have potential ramifications for methods of collecting judgemental data, such as Lickert responses on questionnaires.

In the literature, judgement involving the processing of multiple pieces of information is considered to be serial. As Goldberg (1968) notes "The various studies can thus be viewed as repeated sampling from a uniform universe of judgement tasks involving the diagnosis and predication of human behavior." Researchers assume a measurement model where a decision is made based on the first piece of information and this is sequentially updated by decisions based on the subsequent presentations of information. Researchers have analysed the resulting decisions in terms of 'primacy' effect, where the subject gives more weight to the information that was first presented, and 'recency' effect, where more weighting is given to the information that was presented last. In many calculations on sequential information, such as the calculation of averages or probabilities, the order in which the information is processed does not affect the result of the calculation. In everyday experiences that order often does matter. When meeting someone for the first time, first impressions can colour future interactions. However, the latest weather forecast or stock market report is given more credence for immediate decisions than those in last week's newspaper. In solving experiential problems, such as repairing a stereo player (Tubbs, Gaeth, Levin, and Van Osdol, 1993)or in medical diagnoses (Chapman, Bergus, Gjerde, and Elstein, 1993) subjects seem to give greater weighting to the more recent information. However, subjects seem to preference the primacy effect when the information is ambiguous (Tolcott, Marvin, and Lehner, 1989).

Bastick, T. (1999, July). *Accuracy of same-subject estimates: Are two judgements better than one.* Paper presentated at the 11th European Meeting of the Psychometric, Lueneburg, Germany.

The accuracy of experts' conditional probability decisions has been widely studied, particularly in risk analysis (Kaplan and Garrick, 1981a, 1981b) and in comparing medical diagnoses, where the base-rate likelihood of a disease in the general population greatly affects the accuracy of current probabilistic diagnoses, with Bayesian analysis (Carter, Butler, Rogers, and Holloway, 1993; Gregson, 1993; Kuipers, Moskowitz, and Kassirer, 1988; Meyer and Pauker, 1987; Pozen, D'Agostino, Selker, Sytkowski, and Hood, 1984). For example, when screening for HIV, two patients may show equal likelihood of being positive, but if one patient comes from a population where AIDs is unlikely and the other from a high risk population, these base-rates must be considered to avoid the misery caused by a false positive diagnoses (Meyer and Pauker 1987). However, as Koehler (1993) warns, the base-rates that are usually used for Bayesian analysis are not necessarily the same as prior beliefs. Edwards (1968) found these beliefs considerably influence subject's decisions.

Giving inappropriate consideration to the base-rate is known as the 'base-rate fallacy'. Studies have shown that here are many such problems with the accuracy of human judgement (Kahneman, Slovic, and Tversky, 1982) including adjusting, anchoring, availability, the conjunction fallacy, overconfidence, and representativeness. For example, "The previous review of this field (Slovic, Fischoff & Lichtenstein 1977) described a long list of human judgmental biases, deficiencies, and cognitive illusions. In the intervening period this list has both increased in size and influenced other areas of psychology (Bettman 1979; Mischel 1979; Nisbett & Ross 1980)." Einhorn and Hogarth (1981). Gigerenzer, (1991a, 1991b) has critically evaluated many of these studies. It appears that accuracy and error are linked to intuitive recognition processes (Anderson and Milson, 1989).

In this experiment the above problems are avoided by asking a subject to give two concurrent estimates of the same data. The mean of this double estimate is then calculated and subtracted from the true value so that the its percentage error can be found by dividing this difference by the true value. Independently of this double estimate the subject is also asked to make a single estimate of the same data so that the percentage error of this single estimate can be calculated in the same way. We can then test the assumption that the double estimate is simple a repeated single estimate by comparing the accuracy of the mean double-estimate with the accuracy of the single estimate. For if the double estimate is a repartition of the single estimate process, each of the double estimates will have the same random error about the same mean and so result in equal accuracy. If, however, the mean double-estimate differs in accuracy, then we might infer that a different process is being used. If the mean double-estimate is more accurate then this result would have major implications for collecting judgement data, as for Lickert questionnaire responses.

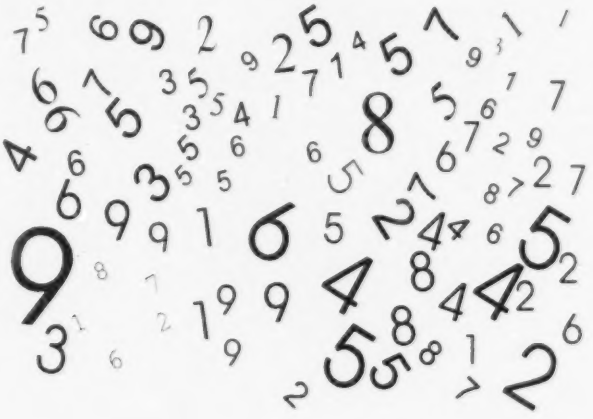
Method

Subjects (n=187) taking an introductory university psychology course competed in speed and accuracy of estimation for ten $10 prizes under extremely server time-limited conditions. The task was to make 72 estimates in 7 minutes. The subjects were presented with 12 sets of numbers randomly positioned on one side of an A4 sheet of paper (11.75"x14.5"). The same 12 sets appeared on the other side of the papers in a disguised form and in a different random order. The 12 sets varied in the amount of numbers they contained (3 difficulty levels - 27, 87 and 146 numbers) and in the range of numbers (4 difficulty levels 0-9, 0-99, 0-999 and 0-9999). The numbers in each set were randomly generated according to these conditions. 'How many number', the 'range' and the total of each of the 12 sets are shown in Table 1. For the 12 sets on one side of the paper, subjects were asked to make single estimates of 'how many numbers' were in each set and the 'total' of the numbers in each set. For the equivalent disguised sets on the other side of the paper subjects were asked to make double estimates of 'how many numbers' and the 'total' of each set.

Table 1: Difficulty levels for the 12 stimulus groups

Random number groups for Single and Double Estimates												
Group	1	2	3	4	5	6	7	8	9	10	11	12
Range	9	99	999	9999	9	99	999	9999	9	99	999	9999
no.	27	27	27	27	87	87	87	87	146	146	146	146
Totals	135	1115	13296	127185	444	4346	49560	415034	720	7387	70970	782378

Figure 1 illustrates the two versions of group 5 with their accompanying instructions. The groups were disguised by randomly changing the position, the font, the size and the rotation of the numbers as well as positioning the groups in a different random order on the other side of the page. Hence, there were two experiments, a 'numbers' task and a 'totals' task. Both experiments were a 3x4 two factor complete randomised block design replicated over the 50 subjects.



			1st estimate	2nd estimate
How many numbers	1A1	**How many numbers**	9B1	9B2
What is their total	1A2	**What is their total**	9B3	9B4

Figure 1: Single estimate and double estimate formats of stimulus group 5

Results

The time conditions were so severe that only 26.7% of the subjects (n=50) completed the tasks. There were 15 males and 35 females aged between 19 and 38 years. Further, (i) 36 subjects gave 358 extremely high or extremely low estimates. (ii) Of these 33 subjects responded with unbalanced extreme estimates. That is, rather than giving consistently high or consistently low estimates, one of their estimates was either extremely high or extremely low. These results indicate the levels of task complexity were too high for the stress level of timed competition used.

Table 2. shows the lack of pattern in the correlation between the single estimates and the mid-double estimates for both the 'numbers' and 'totals' task.

BEST COPY AVAILABLE

Table 2: Correlations of single estimates with mid-double estimates of 'how many numbers and 'totals'

Correlations of single and mid-double estimates of "how many" numbers												
Set No.	1	2	3	4	5	6	7	8	9	10	11	12
Corr	0.1853	0.1438	0.0962	-0.014	0.1077	0.0093	-0.112	0.0538	-0.061	0.0546	0.4081	-0.029
n	50	49	50	50	49	49	50	50	50	50	50	49
sig	P= .198	P= .324	P= .506	P= .926	P= .461	P= .949	P= .440	P= .711	P= .674	P= .706	P= .003	P= .842
Correlations of single and mid-double estimates of "total" of numbers												
Set No.	1	2	3	4	5	6	7	8	9	10	11	12
Corr	0.4383	0.015	0.9382	0.9686	0.0051	0.0844	0.7414	0.4419	-0.023	0.9981	-0.018	-0.024
n	49	49	50	50	49	49	49	49	50	50	50	48
sig	P= .002	P= .918	P= .000	P= .000	P= .972	P= .564	P= .000	P= .001	P= .876	P= .000	P= .904	P= .873

The MANOVA results indicated significant within subject effects for estimates of 'how many numbers' (p=0.000) for the 'totals' (p=0.000) and for the interaction (p=0.000). However, the main factor effects and their interaction was not significant, although the 'how many numbers' factor came close to significance (F=2.95, p=0.063). This again supports the above conclusion that the levels of both factors presented too complex a task at the given level of focused attention, the 'totals' more so than the 'how many numbers'. Like the correlations, the paired t-tests showed no discernable pattern which also supported the above conclusion. The results indicated that the lowest/easiest level of the 'how many numbers' task was the limit of complexity under these conditions. The stimulus materials for this simplest level are shown in Figure 2. The percentage accuracy results for these 4 groups of 27 numbers are shown in Figure 3.

Single estimates 'how many number' = 27

Group 1 range 0-9 Group 2 range 0-99 Group 3 range 0-999 Group 4 range 0-9999

Double estimates 'how many number' = 27

Group 1 range 0-9 Group 2 range 0-99 Group 3 range 0-999 Group 4 range 0-9999

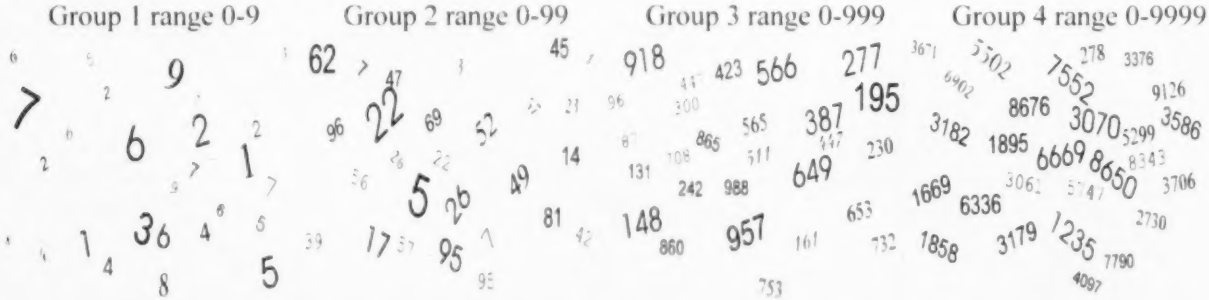


Figure 2. Four stimulus group pairs for the lowest difficult level n=27 of 'How many numbers' task

BEST COPY AVAILABLE

% error in single v mid-double estimate of 'how many numbers'				
Set no.	1	2	3	4
Single	13.93	13.11	20.37	37.11
Mid-double	13.96	18.48	17.78	18.93

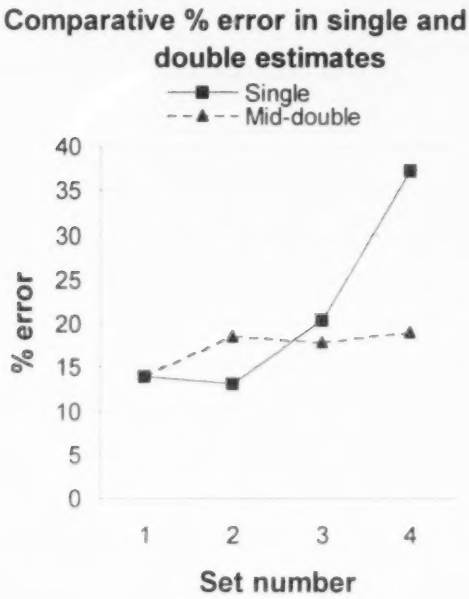


Figure 3. Greater percentage accuracy of the mid-double estimate compared with the traditional single estimate

Conclusion

It is clear from the numbers of unbalanced extreem responses, non-patterened correlations, MANOVA and paired t-tests that all levels of the 'totals' task were too complex for these subjects at this level of timed competition stress and that the numbers task at the lowest levels lowest complexity were at the limit of accurate judgement for these subjects under these conditions. The task at this level demonstrated that the mean of the double-estimate is more accurate than the siggle estimate. This result indicates that the process of double estimate is not a simple replication of the single estimate process. Under conditions where stress and cognitive load are optimised, the mid-double estimate is more accurate than the single estimate.

The greater accuracy produced by this process has wide implications for collecting judgement data. For example, where data is highly valued than it might be more cost effective to collect a double judgement response and use the mean of this double estimate. In practice this might require asking for two Lickert response for each question rather than the traditional single response. Further research is in progress to investigate the improved accuracy of the mid double estimate in contexts of appropiate task complexity.

References

Anderson, J.R., and Milson, R. (1989). Human Memory: an Adaptive Perspective. *Psychological Review, 96,* 783-719.

Carter, B.L., Butler, C.D., Rogers, J.C., and Holloway, R.L. (1993). Evaluation of Physician Decision Making With the Use of Prior Probabilities and a Decision-analysis Model. *Archives of Family Medicine, 2,* 529-534.

Chapman, G.B., Bergus, G.R., Gjerde, C., and Elstein, A.S. (1993). Sources of Error in Reasoning about a Clinical Case: Clinicians as Intuitive Statisticians (Meeting Abstract). *Medical Decision Making, 13,* 382.

Edwards, W. (1968). Conservatism in human information processing. In B. Kleinmuntz (Ed.), *Formal Representation of Human Judgment* (pp. 17-52). New York: Wiley.

Einhorn, H. and Hogarth, R.(1981). Behavioral decision theory: Processes of judgment and choice. *Annual Review of Psychology*, (32), 53-88.

Gigerenzer, G. (1991a). From tools to theories: A heuristic of discovery in cognitive psychology. *Psychology Review, (98),* 254 267.

Gigerenzer, G. (1991b). How to make cognitive illusions disappear: Beyond "Heuristics and Biases, *European Review of Social Psychology, (2),* 83-115.

Goldberg, L. R. (1968). Simple models or simple processes? Some research on clinical judgments. *American Psychologist, 23*(7), 483-496

Gregson, R.A.M. (1993). Which Bayesian Theorem could be compared with real behavior? Commentary on Koehler on base-rate. *Psycoloquy 4*(50)

Hogarth, R.M., & Einhorn, H.J. (1992). Order effects in belief- updating: The belief-adjustment model. *Cognitive Psychology, 24,* 1-55.

Kahneman, D., Slovic, P. and Tversky, A. (eds.) (1982). *Judgment Under Uncertainty: Heuristics and Biases*. Cambridge University, Cambridge, England, pp. 509-520.

Kaplan, S. and Garrick, B. J. (1981a). On the quantitative definition of risk. *Risk Analysis 1*, 11-27.

Kaplan, S. and Garrick, B. J. (1981b). Some misconceptions about misconceptions: A response to Abramson. *Risk Analysis, 1*, 230-233.

Koehler, J.J. (1993). The base rate fallacy myth. *Psycoloquy, 4*(49)

Kuipers, B., Moskowitz, A.J., and Kassirer, J.P. (1988). Critical decisions under uncertainty: Representation and structure. *Cognitive Science, 12*, 177-210.

Meyer, K.B., and Pauker, S.G. (1987). Screening for HIV: Can we afford the false positive rate? *New England Journal of Medicine, 317*, 238-241.

Pozen, M.W., D'Agostino, R.B., Selker, H.P., Sytkowski, P.A., and Hood, W.B., Jr. (1984). A predictive instrument to improve coronary-care-unit admission practices in acute ischemic heart disease: A prospective multicenter clinical trial. *New England Journal of Medicine, 310*, 1273-1278.

Tolcott, M.A., Marvin, F.F., and Lehner, P.E. (1989). Expert decision making in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics, 19*, 606-615.

Tubbs, R.M., Gaeth, G.J., Levin, I.P., and Van Osdol, L.A. (1993). Order effects in belief updating with consistent and inconsistent evidence. *Journal of Behavioral Decision Making, 6*, 257-269. ■

BEST COPY AVAILABLE